



## Sequential pattern mining algorithm for automotive warranty data

Jirachai Buddhakulsomsiri<sup>a,\*</sup>, Armen Zakarian<sup>b</sup>

<sup>a</sup>*School of Manufacturing Systems and Mechanical Engineering, Sirindhorn International Institute of Technology, Thammasat University, Thammasat-Rangsit, P.O. Box 22, Pathumthani 12121, Thailand*

<sup>b</sup>*Department of Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn, Dearborn, MI 48128, USA*

### ARTICLE INFO

#### Article history:

Received 22 June 2007

Received in revised form 31 October 2008

Accepted 3 November 2008

Available online 18 November 2008

#### Keywords:

Data mining

Sequential patterns

Warranty

Association rule

### ABSTRACT

This paper presents a sequential pattern mining algorithm that allows product and quality engineers to extract hidden knowledge from a large automotive warranty database. The algorithm uses the elementary set concept and database manipulation techniques to search for patterns or relationships among occurrences of warranty claims over time. These patterns are represented as IF–THEN sequential rules, where the IF portion of the rule includes one or more occurrences of warranty problems at one time and the THEN portion includes warranty problem(s) that occur at a later time. Once sequential patterns are generated, the algorithm uses rule strength parameters to filter out insignificant patterns, so that only important (significant) rules are reported. Significant patterns provide knowledge of one or more product failures that leads to future product fault(s). The effectiveness of the algorithm is illustrated with the warranty data mining application from the automotive industry. A discussion on the sequential patterns generated by the algorithm and their interpretation for the automotive example are also provided.

© 2008 Elsevier Ltd. All rights reserved.

## 1. Introduction

Many industries, including the automotive industry are faced with the tasks of improving product quality and minimizing warranty costs. Product quality is by-product of the effectiveness of product development processes and their production systems. Thus, product quality can be improved through continuous improvements in product design and development of robust manufacturing and assembly systems. However, no matter how well a product is designed and manufactured, it may fail in the usage environment, either by chance or by some assignable causes. When a product fails within a certain time period, the warranty is a manufacturer's assurance to a buyer that the product will be repaired without a cost to the customer. In a service environment where dealers are more likely to replace than to repair, the cost of component failure during the warranty period can easily equal three to ten times the supplier's unit price (Baird 2000; Feng, Wang, & J. 2001; Cali 1993). Consequently, companies invest significant amounts of time and resources to monitor, document, and analyze product warranty data.

Product quality problems are monitored during the warranty period through the claims filed against the products. This process generates large volumes of warranty data records, such as product problems in the form of repair related labor codes, problem

descriptions, actions taken, repair dates, and repair costs (labor and parts). Sequential pattern analyses of these data records may provide significant benefits to product manufacturers. A sequential pattern analysis searches for patterns or relationships between data objects in a database that occur over time. The analysis is particularly of interest to automotive Original Equipment Manufacturers (OEM), because it identifies important sequential relationships between various product faults. For example, sequential pattern analysis results may reveal a fault pattern that shows how previous product failures may have led to other product fault(s) at a later time. This knowledge enables companies to effectively predict or discover the root causes of failures that are caused by, or are associated with, the earlier problems. This helps in formulating an action plan to remedy the problems and improve product quality, which leads to significant savings in warranty costs and the attainment of product goodwill.

In this paper, a sequential pattern mining algorithm for automotive warranty data is presented. The proposed algorithm is based on the elementary set concept and database manipulation techniques. The algorithm is constructed to search for significant sequential patterns in preprocessed data sets that are obtained from a large automotive warranty database. The sequential patterns are represented in a form of IF–THEN association rules, where the IF portion of the rule includes quality/warranty problems, represented as labor codes, that occurred in an earlier time, and the THEN portion includes labor codes that occurred at a later time. Once a set of unique sequential patterns is generated, the algorithm applies a set of thresholds to evaluate the significance of

\* Corresponding author. Tel.: +66 2986 9009x2109; fax: +66 2986 9112.

E-mail addresses: [jirachai@siit.tu.ac.th](mailto:jirachai@siit.tu.ac.th) (J. Buddhakulsomsiri), [zakarian@umich.edu](mailto:zakarian@umich.edu) (A. Zakarian).

the rules and the rules that pass these thresholds are reported in the solution. The major differences of the proposed approach and those reported in the literature are presented at the end of this section.

Several association rule mining algorithms (Agrawal & Srikant 1994; Agrawal & Shafer 1996; Han & Kamber 2006) and sequential pattern mining algorithms (Agrawal & Srikant 1995; Thomas & Sarawagi 1998; Pei et al. 2004) have been reported in the literature. Agrawal and Srikant (1994) introduced an Apriori algorithm that generates significant association rules between items in a database such that support and confidence of the rules are greater than the user-specified thresholds. However, the algorithm generates a large number of candidate itemsets, whose sizes grow exponentially with the size of a database. To overcome this problem, Agrawal and Srikant (1995) introduced three different Apriori algorithms that define the problem of sequential pattern mining as finding the maximal (longest) sequences of items that have a certain user-specified minimum support. These algorithms use candidate generation technique to address the scalability related shortcomings of their previous approach. Bayardo and Agrawal (1999) proposed metrics for ranking association rules and introduced an algorithm that uses rule support and confidence for extracting best rules from the large data-sets. Pei et al. (2004) proposed the efficient PrefixSpan approach for sequential pattern mining. In PrefixSpan, the global database is projected into a set of smaller (local) databases and sequential patterns are constructed by exploring frequently occurring datasets of local databases.

Many new efficient algorithms are proposed to mine sequential patterns. The differences between these algorithms are mostly related to how they improve computational time by imposing some constraints on the mining process, or in some subtle differences in how they handle the sequential mining process. For example, Yun (2008) uses weight constraints to reduce the number of unimportant patterns, Chen, Cao, Li, and Qian (2008) incorporate user-defined constraints so that the discovered knowledge better meets user needs, Massegia, Poncelet, and Teisseire (2008) introduce time constraints in early stages of the data mining process, and Chen and Huang (2008), Fiot, Laurent, and Teisseire (2007) use fuzzy set techniques and the *K*-means algorithm (Kuo, Chao, & Liu 2009) to achieve better computational efficiency.

Kum, Chang, and Wang (2006) proposed a new sequential pattern mining method based on multiple alignment (rather than the usual support-based approach) for mining multiple databases. Multiple databases are mined and summarized at the local level, and only the summarized patterns are used in the global mining process. Laur, Symphor, Nock, and Poncelet (2007) introduced statistical supports to maximize mining precision and improve the computational efficiency of the incremental mining process. Kum, Chang, and Wang (2007) benchmarked the effectiveness of sequential pattern mining methods by comparing a support-based sequential pattern model with an approximate pattern model based on sequence alignment. Chen and Hu (2007) introduced concepts of recency (an ability to quickly adapt to changes in a database) and compactness, which can cause reasonable time spans for discovering data patterns. They have proposed algorithms that use these concepts to adapt to the frequency of changes in discovered patterns in the database. Lin, Chen, Hao, Chueh, and Chang (2008) introduced the notion of positive and negative sequential patterns, where positive patterns include the presence of an itemset of a pattern, and negative patterns are the ones with the absence of an itemset. Ren, Sun, and Guo (2008) developed an incremental sequential pattern mining process that stores the results from the previous mining and uses them to efficiently mine the database when additional data are added.

Typically, warranty data are strictly confidential for most companies because they relate to product quality, reliability, and are

therefore critical to consumers' product goodwill. As a result, literature on the warranty data analysis of real-life applications is limited to a few published reports (see Blischke & Murthy 1994; Majeske and Herrin 1995, and Lu 1998). Most models and algorithms developed in warranty analysis studies involve warranty cost analysis and can be divided into two categories: (1) one-dimensional studies, which model product failures and warranty costs as a function of the warranty period (see Blischke & Murthy 1996; Sahin & Polatoglu 1998), and (2) two-dimensional studies, which model failures and perform warranty analysis by considering both warranty period and length or frequency of usage (see Murthy, Djameludin, & R.J. 1995; Singpurwalla & Wilson 1998; Majeske 2007). In most studies, the warranty analysis concentrates on: (a) modeling of failure patterns to estimate the number of occurrences (or recurrences) of failures (components, subassemblies, or systems) over the warranty period, assuming all the usage conditions are statistically similar and all the warranty claims are reported with no delay, (b) modeling of rectification costs incurred by failures, and (c) modeling of the expected warranty costs (see Karim, Yamamoto, & Suzuki 2001; Lawless 1998; Polatoglu & Sahin 1998; Suzuki, Yamamoto, Karim, & Wang 2000; Suzuki, Karim, & Wang 2001; Majeske 2007; Fredette & Lawless 2007, and Kulkarni & Resnick 2008). Several studies developed empirical models based on the manufacturer's field data (i.e., failures and costs over the warranty period) for the warranty cost analysis (see Robinson & McDonald 1991; Lawless & Kalbfleisch 1992; Hu & Lawless 1996). Others use probability distribution functions and statistical models for estimating warranty costs with the incomplete data (see Karim et al. 2001; Wang & Suzuki 2001). More recent studies are: Gutiérrez-Pulido, Aguirre-Torres, and Christen (2006), which used a utility-function-based method to determine the appropriate warranty length of a product (brake linings), and Jung and Bai (2007), which applied a bivariate reliability model to estimate the lifetime distribution for products. A comprehensive literature review on warranty data analysis can be found in Murthy and Djameludin (2002).

Although a number of research studies have been reported on warranty analysis, most of them use statistical approaches for cost and/or reliability analysis (Majeske, Lynch, & Herrin 1997; Kalbfleisch, Lawless, & Robinson 1991; Hu & Lawless 1996; Lawless 1998), while very few have applied data mining techniques to warranty data (Hotz et al. 1999, 2001, Buddhakulsomsiri, Siradeghyan, Zakarian, & Li 2006). Hotz et al. (1999) implemented a data mining support environment for planning warranty and goodwill costs in the automotive industry. Regression analysis and back-propagation neural network were used to construct an automatic prediction tool based on the historical warranty data and goodwill costs. Hotz et al. (2001) later developed statistical and machine learning methods for detecting deviation of warranty costs and for the analysis of warranty and goodwill cost statements. Buddhakulsomsiri et al. (2006) implemented a data mining approach to explore the potential benefits of data mining in automotive warranty data analysis. Potential data mining tasks were identified, based on the type of knowledge to be mined. An association rule generation algorithm was developed for important mining tasks. The algorithm was applied to automotive warranty data to illustrate its effectiveness.

In this paper, a new data mining algorithm is presented that uses the elementary set concept of rough set theory (Pawlak 1997) with some important modifications and database manipulation techniques for identifying significant sequential patterns from a large automotive warranty database. Specifically, the algorithm considers all the possible rules that may be generated from a data set rather than the rules determined from the upper and lower approximations of rough set theory. Furthermore, the algorithm proposed in this paper uses important database set operations to

Download English Version:

<https://daneshyari.com/en/article/1135310>

Download Persian Version:

<https://daneshyari.com/article/1135310>

[Daneshyari.com](https://daneshyari.com)